参赛密码 _____

（由组委会填写）

第十届华为杯全国研究生数学建模竞赛

学　　校	中国航天科工集团第二研究院
参赛队号	**83221008**
队员姓名	**1.**崔闪
	2.文毅
	3.张艳欣

参赛密码 _____

（由组委会填写）

第十届华为杯全国研究生数学建模竞赛

题 目 中等收入定位与人口度量模型研究

摘 要：

本文结合武汉大学王祖祥教授提出的洛伦兹曲线构造方法，将基础模型进行组合，自建了两种洛伦兹曲线模型，针对复杂的参数范围，采用变量代换的方法去掉其约束条件。从信号检测与估计理论中得到启发，引入噪声概念求解模型参数。用多项式模型较好的拟合了洛伦兹曲线；对比王祖祥教授提出的十种已有模型，用常用的三种标准判断拟合精度的好坏，定量地说明了自建模型的优点。

近年来国内外对于中等收入人口的比重及意义都有所讨论，也提出了一些测算方法，但都有其局限性，我们建立一种确定的函数关系来克服收入空间法在收入区间取法上的任意性，对于不同地区、不同年份的中等收入人口及其收入区间都有现成可行的划分方法，并改进了人口空间法，方便纵向比较各年中等收入人口与收入。除了对收入空间法和人口空间法的改进，本文还创新地提出了一种判断中等收入人口的指标，并给出了这种方法的经济学意义。

关键词：洛伦兹曲线 噪声概念 多项式模型 中等收入人口 收入空间法 人口空间法

目录

1　问题重述

目前我国的收入分配结构呈现为“金字塔型”。推动我国经济增长尽快转到主要依靠内需特别是消费需求拉动的轨道上来，迫切需要扩大中等收入群体。然而调整国民收入分配结构，就是要构建经济学上所推崇的“橄榄型”收入分配结构，即低收入和高收入相对较少，中等收入占绝大多数的收入分配结构。在这种收入分配格局下，收入差距不大，社会消费旺盛，人民生活水平高，社会稳定。但怎样才能合理地体现出国家的分配格局呢？根据相关文献显示，反映国家的经济规律模型很多，例如洛伦兹曲线模型、多项式模型、密度函数 Kernel 估计法等。

洛伦兹曲线用以比较和分析一个国在不同时代或者不同国家在同一时代的财富不平等该曲线作为一个总结收入和财富分配信息，可以直观的看到一个国家收入分配平等或不平等的状况。洛伦兹曲线模型中 $L(p)$ 表示低于或等于 x 的人口群体拥有收入占总收入的比例，则应有

$$L(p)=\frac{1}{\mu}\int_0^x tf(t)\mathrm{d}t,\quad p=F(x)$$

$L(p)$ 称之为收入分配的洛伦兹曲线。显然，如果 $L_1(p)$ 与 $L_2(p)$ 是两个不同收入分配的洛伦兹曲线，若对任何 $p\in(0,1)$ 都有 $L_1(p)\geq L_2(p)$，则 $L_1(p)$ 对应的收入分配显然更优，因为在 $L_1(p)$ 中，任何低收入端人口拥有的总收入比例更大。图 1 为某收入的密度函数 $f(x)$，记对应的分布函数为 $F(x)$，则 $p=F(x)$ 表示收入低于或等于 x 的人口比例。由于 $F(m)=1/2$，意味着收入大于或等于平均收入的人口一定不到半数，因此是少数。

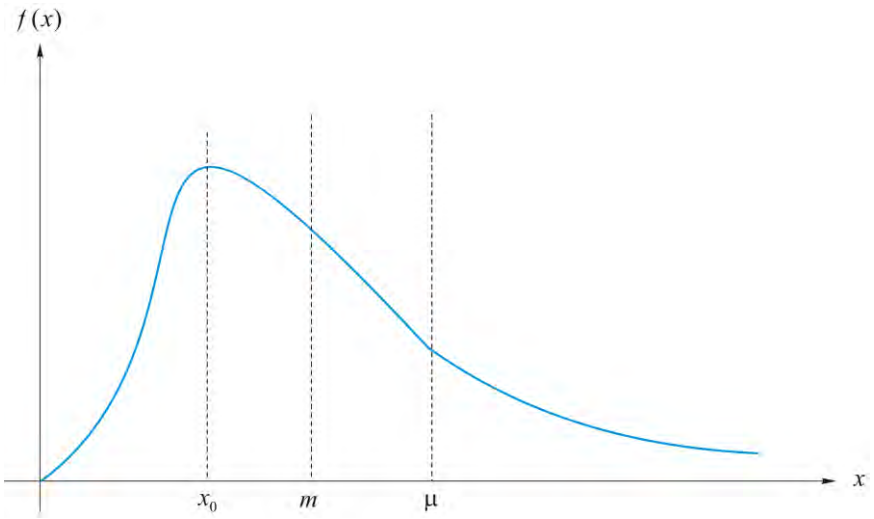


图1 密度函数曲线

图 2 中的红色线表示洛伦兹曲线其中横轴表示人口比例，纵轴表示总收入比例。显然，图中曲线位置越高，所代表的收入分配越平等。

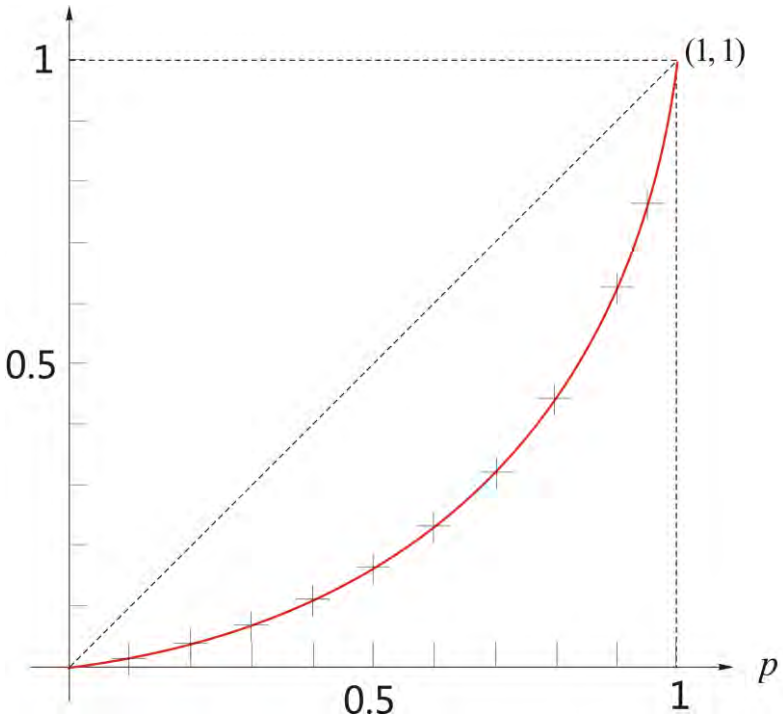


图2 洛伦兹曲线

根据上述对洛伦兹曲线的描述我们需要构造一个能够更好地反映国民收入在国民之间分配的数学模型。而且近年来国内媒体对于扩大中等收入者的比重的意义，扩大中等收入者比重的途径，都有所讨论，看法相对比较一致，但对中等收入者的概念、界定方法或划分标准，如何合理界定其收入的上下限，各种看法出入很多。中等收入者在国家收入分配上起到很大的作用。目前提出的中等收入人群的划分方法仍有一些缺陷，例如由于经济进步，通货膨胀等因素的影响，收入区间是变化的，更多的情形是所有人口的收入都提高了，即全社会的收入区间右移，可见收入区间的范围存在任意性，这就使总想比较各年的中等收入人口出现困难。

因此需要一种很好的模型来改善收入空间法的收入区间的任意性，以及测算出中等收入人口比例大小。我们认为，中等收入人口不单纯是经济学概念，它具有经济、历史、地域和社会等多重规定，目前对中等收入标准的界定，还没有统一的标准。因此找出一种合理的方法来衡量我国的中等收入范围以及中等收入人口比例是十分重要的。本文针对上述情况，通过对已有的数据进行分析建模主要解决以下问题：

- 如何很好的描述我国的国民收入问题；
- 如何才能很好地确定我国中等收入范围和中等收入人群
- 对中等收入人口的变化情况进行合理的分析；
- 对中等收入人口进行重新定义、原理及经济学原理

2　问题分析及符号说明

2.1问题分析

要建立洛伦兹曲线较好的拟合收入分配分组数据，曲线的模型构建显得尤为重要，洛伦兹曲线模型有其固有的基础模型，在自建模型的过程中可以采用多种基础模型进行组合，在这个过程中要注意模型参数的变化，参数的范围很可能会变得相当复杂，不妨采用参数代换的方法去掉约束条件。

自建模型之后，从图像上直观感受其拟合精度，并采取通用的标准与已有的模型进行比较，定量描述模型的好坏。

构建曲线模型是为了将收入分组数据连续化，更好的分析该地区收入分配水平，而确定中等收入以及中等收入人口的范围需要采用特定的方法，已有的收入空间法和人口空间法都存在着一定的缺陷，我们要建立一种确定的函数关系来克服收入空间法在收入区间取法上的任意性，这样对于不同地区、不同年份的中等收入人口及其收入区间都有现成可行的划分方法；我们也可以找到确定的函数来对人口空间法的上下限取值进行估计，这样可以方便的纵向比较各年中等收入人口与收入。

对于中等收入人口的测算方法，除了改进已有的方法之外，还可以重新构建一种指数来描述其在社会总人口中所占的比重，这种指数我们可以在已经大量使用的洛伦兹曲线中找到灵感，用来描述曲线的弯曲程度，借此得到中等收入人口的比重。

2.2符号说明

$f(x)$	收入分配的密度函数	q_h	人口比为 10%的最低收入人口的平均收入
p	收入低于或等于 x 的人口比例	Δq	收入水平差
$L(p)$	低于或等于 x 的人口收入占总收入的比例	φ	收入因子
m	收入中位数点	δ	比例系数
μ	平均收入	Δw	财富比例变化值
x_l	中等收入人口收入下限	ε	财富分布平等指数
x_h	中等收入人口收入上限	R_{mid}	中等收入人口比重
q_l	人口比为 10%的最高收入人口的平均收入		

3　模型的建立与求解

3.1 自建洛伦兹模型一

我们已经知道洛伦兹曲线模型是一种行之有效的方法，我们从大量参考文献中总结出洛伦兹曲线模型的特点，并尝试根据这些特点自己构造洛伦兹模型。洛伦兹曲线模型中的项大部分是由以下五种基础模型组合而成。因此，我们建立模型的时候，也是基于这些基础模型。

$$L_1(p)=p \tag{3-1-1}$$

$$L_2(p)=1-(1-p)^{\beta} \quad \beta\in(0,1] \tag{3-1-2}$$

$$L_\lambda(p)=\frac{e^{\lambda p}-1}{e^{\lambda}-1} \qquad \lambda>0 \tag{3-1-3}$$

$$L_3(p)=1-L_{\lambda_1}(1-p)^{\beta_1} \qquad \beta_1\in(0,1],\lambda_1\in(-\infty,0)\cup(0,\ln\beta_1^{-1}] \tag{3-1-4}$$

$$L_4(p)=1-\left(1-L_{\lambda_2}(p)^{\beta_2}\right) \quad \beta_2\in(0,1],\lambda_2\in[\ln\beta_2,0)\cup(0.+\infty) \tag{3-1-5}$$

基于上述思想，我们首先根据洛伦兹曲线的特点 $L(0,\tau)=0$，确定模型中的常数项为 0，为模型构建 p^{α} 这一项确保在 $p=0$ 的时候 $L(p,\tau)=0$；在根据 $L(1,\tau)=1$，为模型构建 $(1-p)^{\beta}$ 这一项确保在 $p=1$ 的时候 $(1-p)^{\beta}=0$，由于洛伦兹曲线是一个上凸的曲线，在 $p\to1$ 时 $(1-p)^{\beta}\to0$，因此可以看出此项前应加一个负号，并根据 $L(1,\tau)=1$ 加上 1，这样就能保证洛伦兹模型满足条件 $L(0,\tau)=0$ 和 $L(1,\tau)=1$。

下面考虑上述基本模型中的 3-1-3 模型，可以将之加成到 $(1-p)^{\beta}$ 的项中，因为在众多参考文献中，并没有发现 $L_\lambda(p)$ 和 $(1-p)^{\beta}$ 做乘积的模型出现，我们先试验性的列出一种公式，根据公式的曲线特性做出修改，构造出一种新的洛伦兹曲线模型：

$$L(p)=p^{\alpha}[1-L_\lambda(p)(1-p)^{\beta}]^{\nu} \tag{3-1-6}$$

接下来我们确定模型中参数的取值范围，我们根据洛伦兹曲线模型的特点

$$L(0,\tau)=0,\ L(1,\tau)=1,\ L'(p,\tau)\ge0,\ L''(p,\tau)\ge0, \tag{3-1-7}$$

对 3-2-6 求一阶导数和二阶导数，我们可以确定 $L(p)$ 中参数的取值范围：

$$\alpha\ge0,\ \nu\ge0,\ \alpha+\nu\ge1,\ \beta\in(0,1],\ \lambda\in(-\infty,0)\cup(0,\ln\beta^{-1}] \tag{3-1-8}$$

对于一些结构较为复杂、变元较多的数学问题 ，引入一些新的变量进行代换，以简化其结构，从而达到解决问题的目的这种方法叫做变量代换法。变量代换法是一种非常有效的解题方法，尤其是处理一些复杂的不等式问题，效果明显。

合理代换往往能简化题目的信息，凸显隐含条件，沟通量于量之间的关系，对发现解题思想，优化解题过程有着重要的作用。

在本模型中 $\beta \in (0,1]$ 很容易想到利用三角函数，故此采用 $\sin^2\theta = \beta$；$\alpha \geq 0$ 则联系到 $e^{\eta} = \alpha$，这样就将约束条件转变为无约束条件。

然后通过运用 lsqcurvefit()函数，并根据表 7 的数据得到 $\alpha, \beta, \nu, \lambda$ 这四个未知量的估计值：

$\alpha = 2.0424,\ \beta = 0.4019,\ \nu = 0.2403,\ \lambda = -31.6917$

通过 matlab 程序得到洛伦兹曲线的图形：

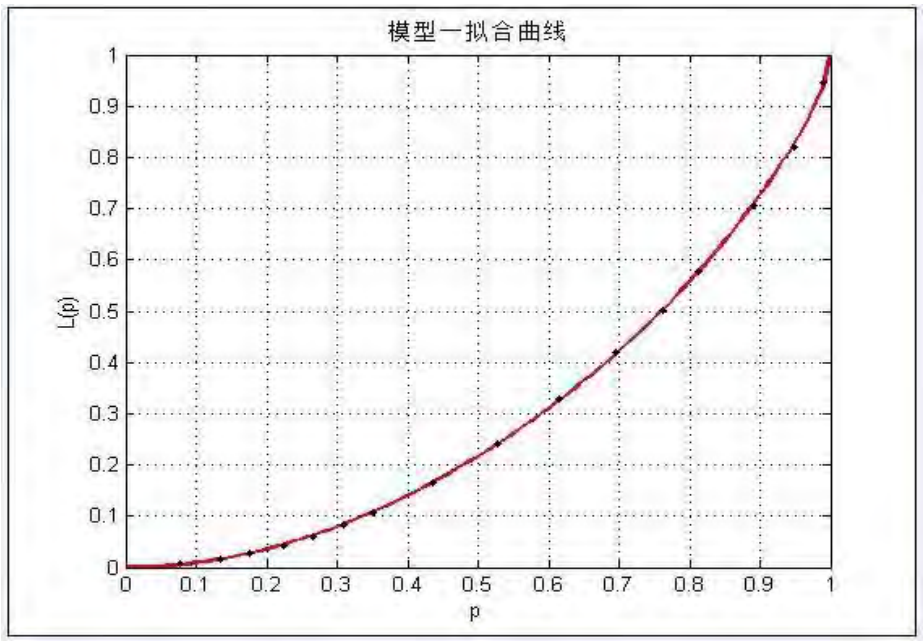


图3 自建洛伦兹模型一拟合曲线图

表1　自建洛伦兹模型一误差特性

	均方误差（MSE）	平均绝对误差（MAE）	最大绝对误差（MAS）
自建洛伦兹模型一	3.06365×10^{-6}	0.0014	0.0232

3.2 自建洛伦兹模型二

另外一种构建洛伦兹模型的思路是将弓形曲线 $L_1(p)=Ap^{\alpha}(1-p)^{\beta}$ 与(0,0),(1,1)的对角线 $L_2(p)=p$ 结合合起来，从而得到一种洛伦兹曲线模型：

$$L(p)=p-Ap^{\alpha}(1-p)^{\beta} \qquad 3\text{-}2\text{-}1$$

式中，A, α, β 是参数，$p = F(x)$ 表示收入低于或等于 x 的人口比例，收入低于或等于 x 的人口人口拥有收入占总收入的比例为 $L(p)$。

由洛伦兹曲线的四个约束条件，$L(0,\tau)=0$，$L(1,\tau)=1$，$L'(p,\tau)\geq 0$，$L''(p,\tau)\geq 0$对 3-3-1 求一阶和二阶导数：

$$\begin{aligned}\frac{dL(p)}{dp}&=1-A\left[\alpha p^{\alpha-1}(1-p)^{\beta}-\beta p^{\alpha}(1-p)^{\beta-1}\right]\\&=1-\left(\frac{\alpha}{p}-\frac{\beta}{1-p}\right)Ap^{\alpha}(1-p)^{\beta}\\&=1-\left(\frac{\alpha}{p}-\frac{\beta}{1-p}\right)\left(p-L(p)\right)\\&=1-\alpha\frac{p-L(p)}{p}+\beta\frac{p-L(p)}{1-p}\end{aligned}$$

$$\frac{d^2L(p)}{dp^2}=\left(p-L(p)\right)\left[\frac{\alpha(1-\alpha)}{p^2}+\frac{\beta(1-\beta)}{(1-p)^2}+\frac{2\alpha\beta}{p(1-p)}\right]$$

满足一阶和二阶导数同时大于零的参数范围是：$A>0,\alpha,\beta\in(0,1)$。

下面重点说明对于参数求解的另一种不同的方法：在学习信号检测与估计理论的过程中使用过噪声的概念，本组成员讨论后认为可以借鉴使用，下面介绍求解参数的方法。

首先对提出的洛伦兹曲线进行变换：

$$L(p)=p-Ap^{\alpha}(1-p)^{\beta}$$

$$\Downarrow$$

$$\begin{aligned}\ln\left(p-L(p)\right)&=\ln A+\alpha\ln p+\beta\ln(1-p)\\&=\left[1,\ln p,\ln(1-p)\right]\left[\ln A,\alpha,\beta\right]^T\end{aligned}$$

引入噪声的概念之后，加入噪声$n(k)$，并令

$Z(k)=\ln\left(p-L(p)\right),V(k)=\left[1,\ln p,\ln(1-p)\right],a(k)=\left[\ln A,\alpha,\beta\right]^T$，上式变形为：

$$Z(k)=V(k)a(k)+n(k)$$

根据最小二乘算法的定义，使误差平方和，也就是这里所说的噪声的平方和最小，

$$\|\delta\|_2^2=\sum_{k=0}^{N}\delta_i^2=\min\sum_{k=0}^{N}n(k)^2=\min\sum_{k=0}^{N}\left[Z(k)-V(k)a(k)\right]^2$$

为简化书写，同时也为了方便阅读，这里不妨做如下简写：

$Z(k)=z,\ V(k)=v,\ a(k)=a,\ n(k)=n$，式中$z,v,a,n$都是向量形式。

因此有：

$$
\begin{aligned}
n^T n &= \left(z - va\right)^T \left(z - va\right) \\
&= \left(z^T - a^T v^T\right)\left(z - va\right) \\
&= z^T z - z^T va - a^T v^T z + a^T v^T va \\
&= \left(a - \left(v^T v\right)^{-1} v^T z\right)^T v^T v \left(a - \left(v^T v\right)^{-1} v^T z\right)
\end{aligned}
$$

上式可以清楚地看出，要使噪声的平方和最小，$a = \left(v^T v\right)^{-1} v^T z$。

下面根据表 7 中的数据（去掉最后一行进行处理），得到v是一个16×3的矩阵，z是一个16×1的矩阵，计算得到的a是一个3×1的矩阵，上文中也提到了$a(k) = \left[\ln A, \alpha, \beta\right]^T$，这样通过矩阵的乘法运算，并将$\ln A$的数值做指数运算得到$A$,这样的一组参数值为：

$$A = 0.8417,\ \alpha = 0.9328,\ \beta = 0.6349$$

将上述参数带入洛伦兹曲线模型二中，用 matlab 绘出图形为：

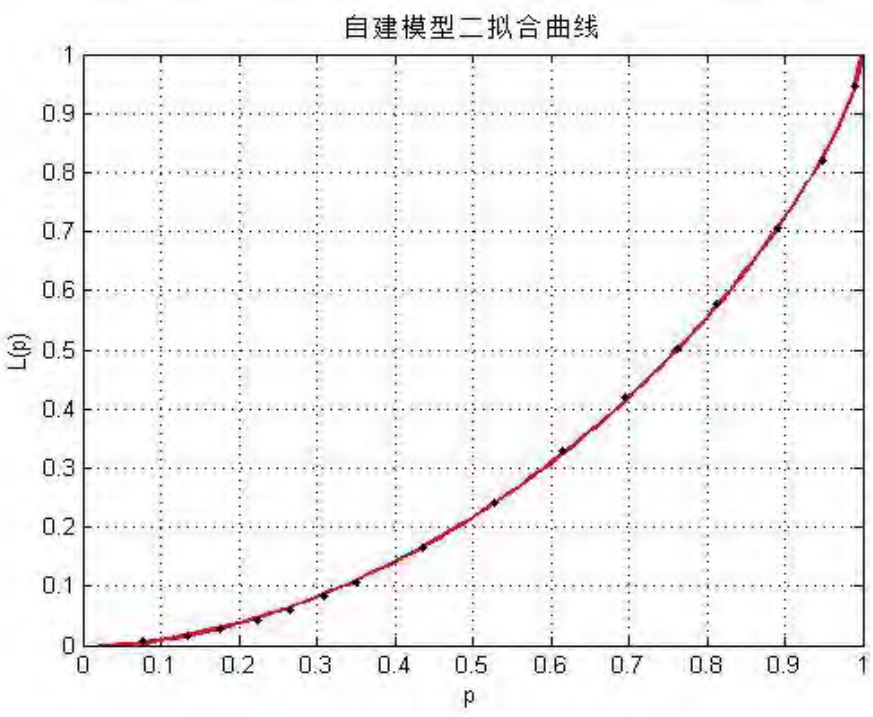


图4 自建洛伦兹模型二拟合曲线图

从图中可以看出，绘出的曲线与数据点的拟合程度还是很好的，从精度计算中得到结果也令我们满意。这样除了使用 lsqcurvefit()函数方法直接用 matlab 求出参数这一种方法之外，参照信号检测与估计的方法，引入噪声的概念，一样可以求出洛伦兹曲线模型的参数。

下面从定量的角度分析，

表2　自建洛伦兹模型二误差特性

	均方误差（MSE）	平均绝对误差（MAE）	最大绝对误差（MAS）
自建洛伦兹模型二	8.2390×10^{-6}	0.0025	0.0402

从以上三种标准中可以看出，我们采用的噪声概念求参数方法得到的参数，应用在提出的洛伦兹模型二上，三种精度标准与洛伦兹曲线模型一都处于同一数量级上，可见引入噪声概念后，这种求参方法是切实可行的。

3.3 多项式模型

在洛伦兹曲线拟合方面，题目中提到使用洛伦兹曲线模型是比较理想的方法之一，但并不排除通过多项式模型、样条函数模型逼近来确定的洛伦兹曲线，因此，我们尝试使用多项式模型。因为多项式模型结构比较简单，我们知道一次项的多项式是一条直线，二次项的多项式是一条抛物线，我们根据洛伦兹曲线的特性：

$$L(0,\tau)=0,\quad L(1,\tau)=1,\quad L'(p,\tau)\geq 0,\quad L''(p,\tau)\geq 0 \qquad 3\text{-}3\text{-}1$$

即 $L(0,\tau)$ 在[0,1]上是凸增函数，构建出多项式的基本模型为：

$$L(p)=p+(1-p)[\mathrm{a}_1\,p^n+a_2p^{n-1}+a_3p^{n-2}+\cdots+a_np] \qquad 3\text{-}3\text{-}2$$

我们通过比较以及误差分析可知当 n=3 时多项式的洛伦兹曲线比较符合标准。因此我们可以得到的多项式为：

$$L(p)=p+(1-p)[\mathrm{A}\,p^3+\mathrm{B}\,p^2+\mathrm{C}\,p] \qquad 3\text{-}3\text{-}3$$

其中 A,B,C 是未知参数，拟合附录中表一的数据，使用非线性最小二乘法求解

$$\min\sum_{i=1}^{n}(L(p_i)-L_i)^2 \qquad 3\text{-}3\text{-}4$$

在 matlab 中使用 lsqcurvefit 函数进行非线性最小二乘拟合得到多项式中未知的参数，从而得出 $L(p)$ 的表达式，在这里，我们将表 7 中的数据输入到 matlab，通过拟合，得到参数 A，B，C 的值分别为-2.1020　1.2497　-1.2155，将之代入多项式模型，作出的拟合洛伦兹曲线如下图所示，图中标记的点为样本点。经过 matlab 仿真可以得到多项式模型的洛伦兹图形为：

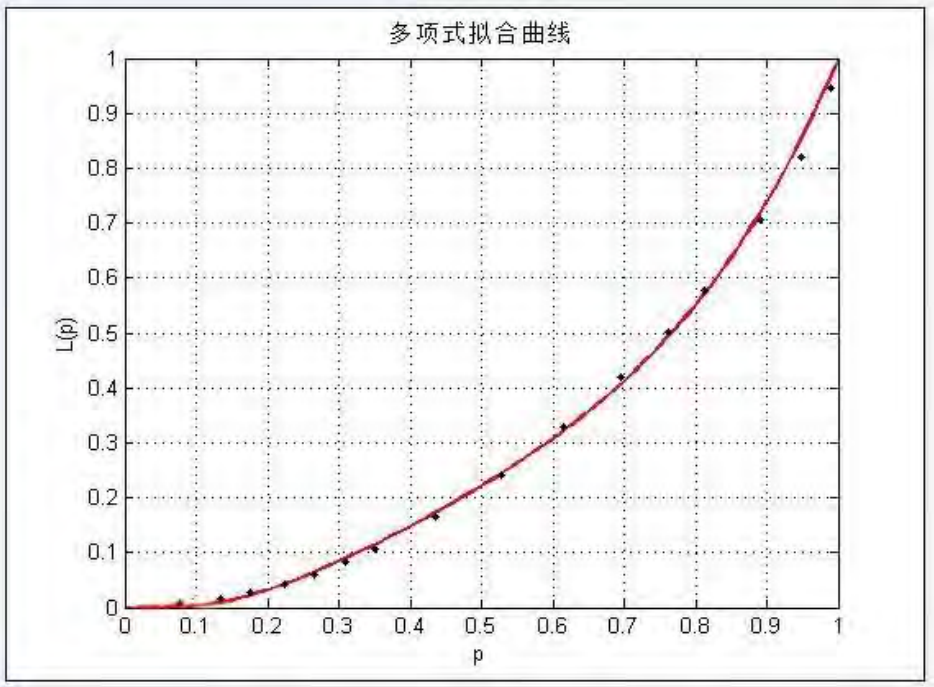


图5 多项式模型拟合曲线图

拟合精度的好坏我们采用以下三种标准进行比较：

① 均方误差（MSE，mean squared error）

$$\frac{1}{n}\sum_{i=1}^{n}\left[L(p_i,\hat{\tau})-L_i\right]^2 \qquad 3\text{-}3\text{-}5$$

② 均绝对误差（MAE,mean absolute error）

$$\frac{1}{n}\sum_{i=1}^{n}\left|L(p_i,\hat{\tau})-L_i\right| \qquad 3\text{-}3\text{-}6$$

③ 大绝对误差（MAS，maximum absolute error）

$$\max_{1\le i\le n}\left|L(p_i,\hat{\tau})-L_i\right| \qquad 3\text{-}3\text{-}7$$

通过运用 matlab 可以获得其拟合精度，其结果如下表：

表3 多项式模型误差特性

	均方误差（MSE）	平均绝对误差（MAE）	最大绝对误差（MAS）
多项式模型	1.5179×10^{-4}	0.0023	0.0376

3.4 多模型对比

从参考文献中，我们找出 10 种洛伦兹模型，如下所示：

$$L_1(p)=L_\lambda(p)=\frac{e^{\lambda p}-1}{e^{\lambda}-1} \qquad \lambda>0 \tag{3-4-1}$$

$$L_2(p)=\left(1-(1-p)^{\beta}\right)^{\nu} \qquad \beta\in(0,1],\nu\geq 1 \tag{3-4-2}$$

$$L_3(p)=p^{\alpha}\left[1-(1-p)^{\beta}\right]^{\nu} \quad \beta\in(0,1],\alpha\geq 0,\nu\geq 1,\alpha+\nu\geq 1 \tag{3-4-3}$$

$$L_4(p)=p^{\alpha}L_\lambda(p)^{\nu} \quad \lambda>0,\alpha\geq 0,\nu\geq 1,\alpha+\nu\geq 1 \tag{3-4-4}$$

$$L_5(p)=p^{\alpha}\left[1-L_\lambda(1-p)^{\beta}\right]^{\nu} \quad \beta\in(0,1],\lambda\in(-\infty,0)\cup\left(0,\ln\beta^{-1}\right],\ \alpha\geq 0,\nu\geq 1/2,\alpha+\nu\geq 1 \tag{3-4-5}$$

$$L_6(p)=p^{\alpha}\left[1-\left(1-L_\lambda(p)\right)^{\beta}\right]^{\nu} \quad \beta\in(0,1],\lambda\in[\ln\beta,0)\cup(0,+\infty),\alpha\geq 0,\nu\geq 0,\alpha+\nu\geq 1 \tag{3-4-6}$$

$$L_7(p)=p^{\alpha}\left[1-L_{\lambda_1}(1-p)^{\beta_1}\right]^{\alpha_1}\left[1-\left(1-L_{\lambda_2}(p)\right)^{\beta_2}\right]^{\nu} \quad \alpha\geq 0,\nu\geq 0,\alpha+\nu\geq 1,\alpha_1\geq 0,\alpha+\alpha_1\geq 1,\alpha_1+\nu\geq 1 \tag{3-4-7}$$

$$L_8(p)=\left(1-(1-p)^{\beta_1}\right)^{\alpha}\left(1-(1-p)^{\beta}\right)^{\nu} \quad \alpha\geq 0,\nu\geq 0,\alpha+\nu\geq 1,\beta\in(0,1],\beta_1\in(0,1] \tag{3-4-8}$$

$$L_9(p)=p^{\alpha}\left(1-(1-p)^{\beta}\right) \qquad \alpha\geq 0,\beta\in(0,1] \tag{3-4-9}$$

$$L_{10}(p)=\delta p^{\alpha}\left[1-(1-p)^{\beta}\right]+(1-\delta)\left[1-(1-p)^{\beta_1}\right]^{\nu} \quad \alpha\geq 0,\nu\geq 0,\beta\in(0,1],\beta_1\in(0,1],\delta\in(0,1) \tag{3-4-10}$$

由于上述模型都是非线性函数，非线性最小二乘估计需要估算参数的数值。当参数范围的结构较为复杂时，我们可以采用参数变换来对约束条件进行变换，这种变换使我们可以采用不受约束的非线性最小二乘估计，这种方法比受约束的非线性最小二乘估计要有效率的多。因此我们用 e^x 代替 3-5-1 中的 λ 即：$\lambda=e^x$ 因此 $L_1(p)$ 变为 $L_1(p)=L_\lambda(p)=\dfrac{e^{e^x p}-1}{e^{e^x}-1}$，并用其替代上述公式中的 $L_\lambda(p)$。

对于每一个模型，我们都是用 matlab 中的非线性最小二乘法拟合得到其未知参数，得到未知参数如下表所示:

表4 各模型参数表

	α	β	v	x	α_1	x_1	x_2	β_1	β_2	δ
$L_1(p)$				0.990						
$L_2(p)$		0.780	1.759							
$L_3(p)$	1.405	0.509	0.465							
$L_4(p)$	0.982		0.477	1.234						
$L_5(p)$	0.039	0.682	1.730	0.062						
$L_6(p)$	0.282	0.730		0.900						
$L_7(p)$	0.725		0.491		0.698	1.254	0.567	0.585	0.702	
$L_8(p)$	1.759	0.200	0.010					0.780		
$L_9(p)$	0.789	0.677								
$L_{10}(p)$	0.799	0.677	1					0.500		0.6

有了参数就可以代入相应的模型中，对其进行 MATALB 仿真得出了它们的洛伦兹曲线以及它们的均方误差、平均绝对误差、最大绝对误差如下表所示：

表5 各模型误差系数

	均方误差（MSE）	平均绝对误差（MAE）	最大绝对误差（MAS）
$L_1(p)$	3.0137×10^{-4}	0.0151	0.2419
$L_2(p)$	2.1740×10^{-5}	0.0042	0.0670
$L_3(p)$	4.40126×10^{-6}	0.0018	0.0283
$L_4(p)$	3.0277×10^{-4}	0.0117	0.1872
$L_5(p)$	6.5897×10^{-7}	6.5411×10^{-4}	0.0103
$L_6(p)$	2.1974×10^{-5}	0.0034	0.0539
$L_7(p)$	3.8670×10^{-5}	0.0050	0.0798
$L_8(p)$	1.1657×10^{-5}	0.0031	0.0492
$L_9(p)$	1.1567×10^{-5}	0.0040	0.0532
自建洛伦兹模型一	3.06365×10^{-6}	0.0014	0.0232

自建洛伦兹模型二	8.2390×10^{-6}	0. 0025	0. 0402
多项式模型	1.5179×10^{-4}	0.0023	0. 0376

下面给出 10 种模型的拟合曲线

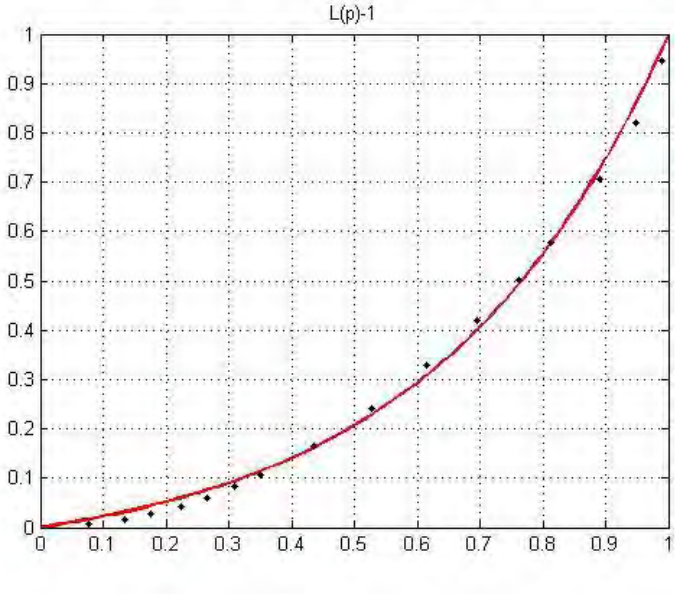


图6 模型一拟合曲线

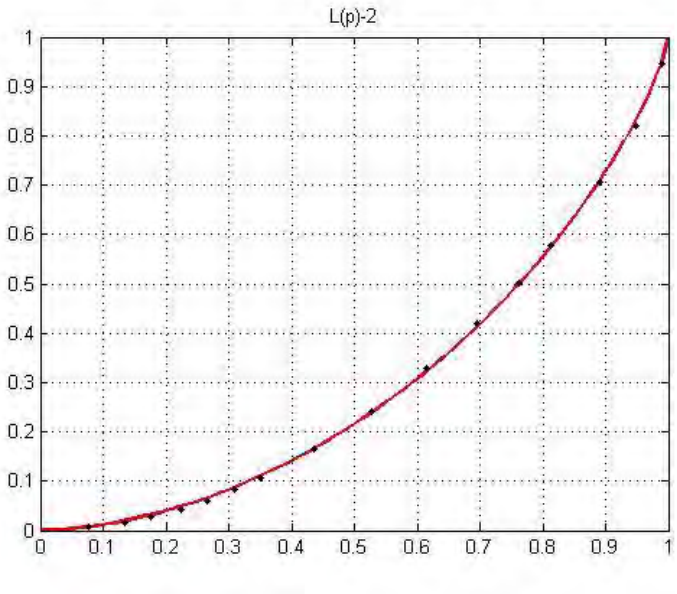


图7 模型二拟合曲线

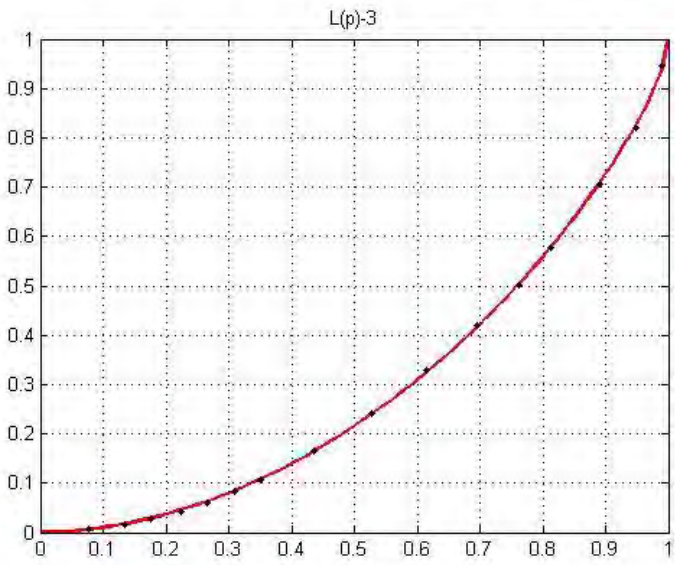


图8 模型三拟合曲线

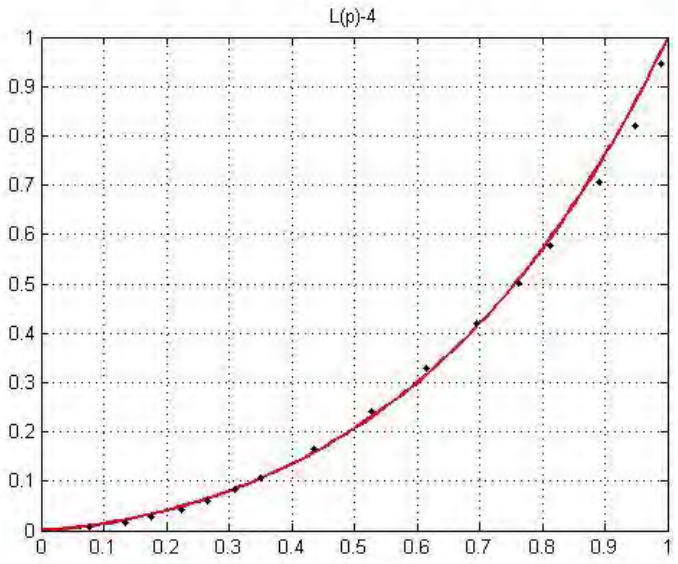


图9 模型拟合四曲线

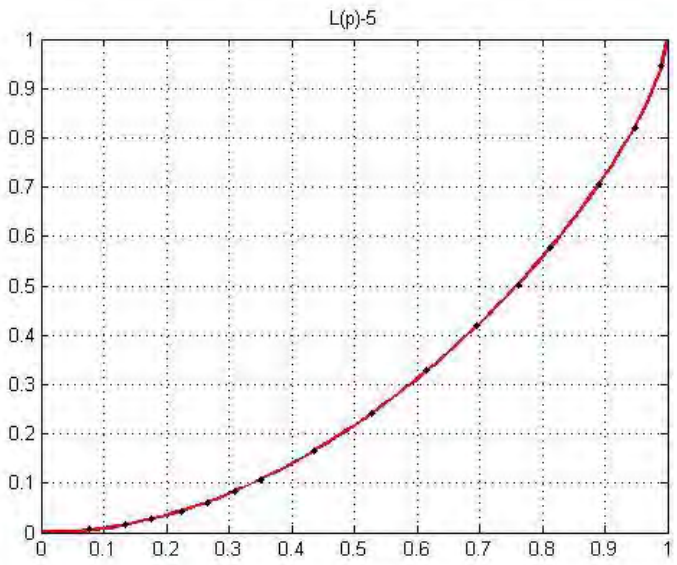


图10　模型五拟合曲线

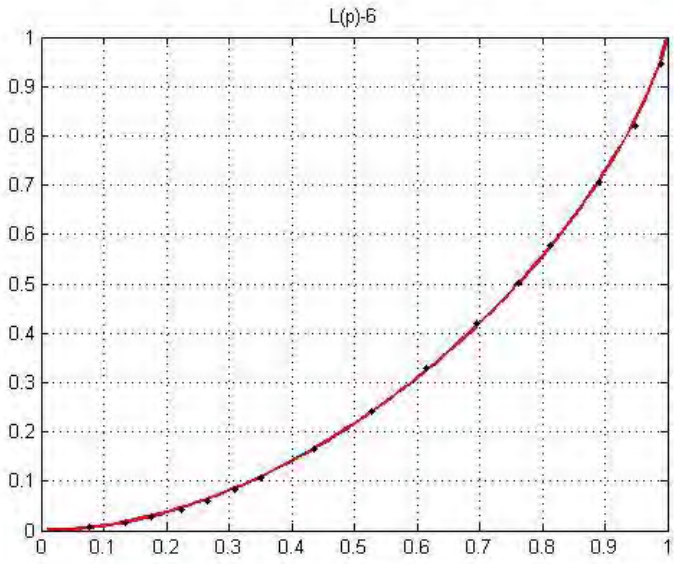


图11　模型六拟合曲线

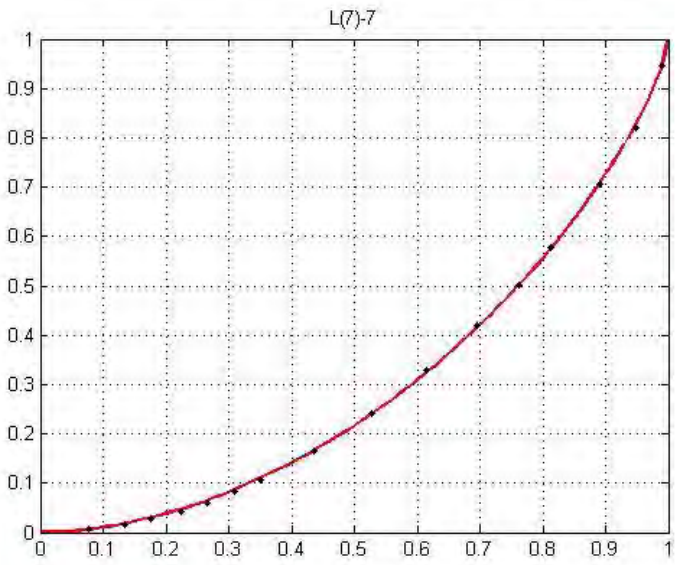


图12　模型七拟合曲线

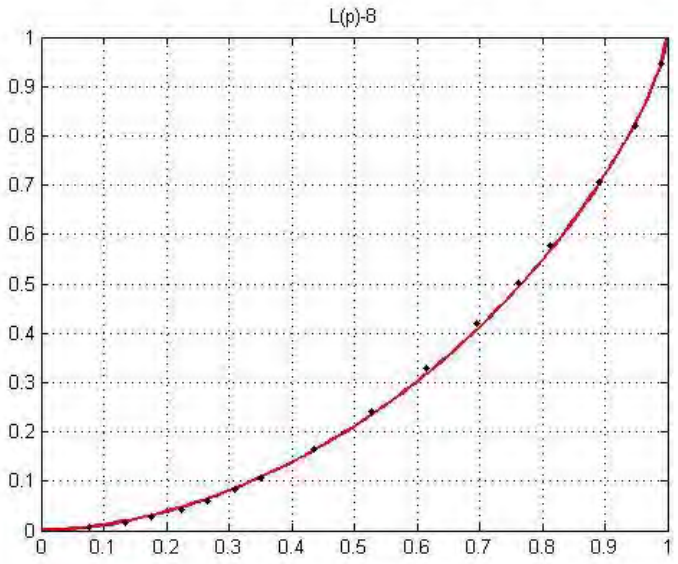


图13　模型八拟合曲线

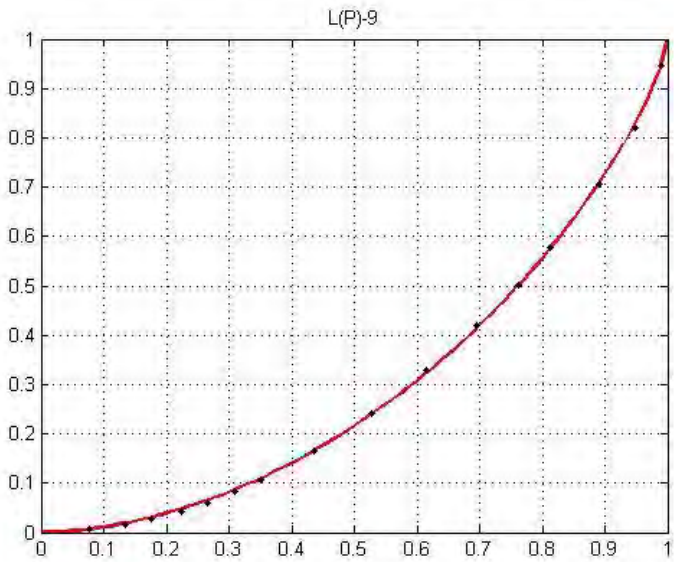


图14　模型九拟合曲线

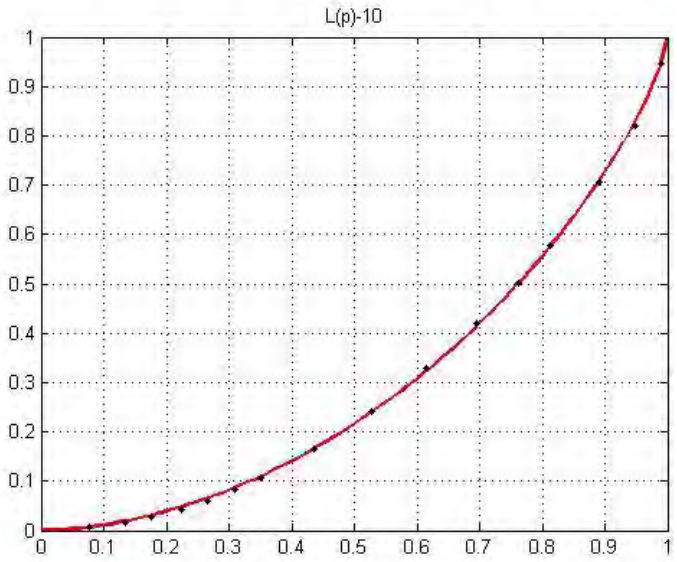


图15　模型十拟合曲线

通过将本文中新设计出的模型：多项式模型、洛伦兹曲线模型一、洛伦兹曲线模型二以及从文献中找出的十种模型的均方误差、平均绝对误差、最大绝对误差进行对比，可以得出：

1、我们所构造出的多项式和洛伦兹模型的均方误差，和文献中找到的十种模型的均方误差都很小，均在在10^{-4}以下，其中洛伦兹曲线模型一的均方误差值在10^{-6}数量级上，比上述十种模型中的一些还要好；

2、平均绝对误差方面，所构建的模型和文献中摘取的十种模型几乎都在一个数量级上，差别不是很大；

3、最大绝对误差的范围在$(0.01,\ 0.25)$之间，我们构建模型的最大误差分别为 0.0232、0.0376，接近于这个范围的下限，精度较高。

4、特别要说明的是模型⑤$L_5(p)=p^{\alpha}\left[1-L_{\lambda}\left(1-p\right)^{\beta}\right]^{\nu}$的均方误差要比其他模型低 1-3 个数量级，平均绝对误差要低 1-2 个数量级，最大绝对误差也要低，因此在上述的十三种模型中，模型⑤$L_5(p)=p^{\alpha}\left[1-L_{\lambda}\left(1-p\right)^{\beta}\right]^{\nu}$的精确度最高，但不可否认的是，我们构建的模型在这三种指标上并不比文献中的已有模型差，反而在有些方面精度更高。

因此本文所设计的洛伦兹模型二比文献中给出的其他模型，能够更好的描述国民收入在国民之间的分配问题。

洛伦兹曲线模型与密度函数 Kernel 估计法

Kernel 估计法即核密度估计，它是用来估计收入分布密度函数的非参数检验方法之一。该方法是基于给定的核函数来推算样本的密度函数，从而找出其分布状态假如寻求一个随机变量 x 的概率密度函数 f(x) 。这里对于 X 的分布不强加任何形式的函数假设其估计量用 $\hat{f}(\mathrm{x})$ 表示。在特定的假设条件下，可以推导出“核密度估计量”的形式如下: $f(\mathrm{x})=\frac{1}{nh}\sum_{i=1}^{\hat{n}}K(\frac{X_i-x}{h})$ 要使用上式，关键问题是选择合适的核函数$K(\cdot)$ 以及最优带宽h ，但核密度函数估计对数据要求较高，尤其是需要大样本的住户调查样本资料，核密度估计在估计边界区域的时候会出现边界效应。

4　中等收入人群的确定

4.1 基于收入空间法的改进

经济理论界考虑取收入落在中位收入m的一个范围内的人口为中等收入人口，可以视这种方法为“收入空间法”，如图 3 所示。取其中收入属于(x_l,x_h)中的人口为中等收入人口，这时中等收入人口比例M显然等于$F(x_h)-F(x_l)$。显然，这种方法中x_l与x_h的取法具有任意性，由于经济进步，通货膨胀等因素的影响，收入的区间是变化的，更多的情形是所有人口的收入都提高了，即全社会的收入区间右移，可见x_l与x_h的任意性使纵向比较各年的中等收入人口时出现困难。

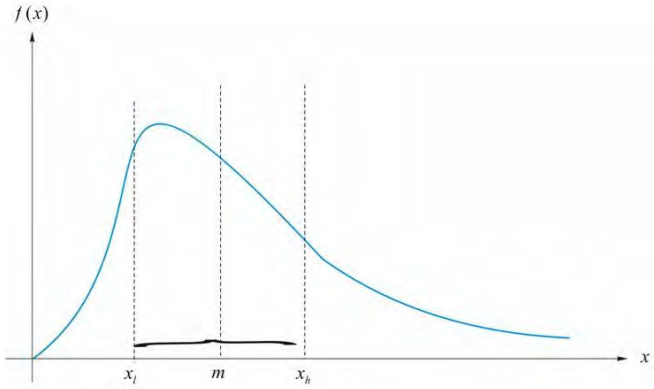


图16　依据中位数确定区间示意图

为了解决中等收入区间取法的任意性，克服由于经济进步，通货膨胀等因素引起的收入区间右移而导致纵向对比困难，我们提出了一种改进的中位数确定法。

中等收入区间应在中位数 m 附近，在收入数轴上向左向右取出一个合理的区间 (x_l, x_h)，收入下限与中位数的差值 Δx_l 和收入上限与中位数的差值 Δx_h 应当综合考虑所有人口的收入水平。通过分析统计数据我们发现，Δx_l 、Δx_h 的取值和整体收入水平呈一个正相关关系，我们取最高收入水平和最低收入水平的插值作为整体收入水平。暂且将这种方法命名为：中位数调整法。

表6　相关参数表

最高收入水平 q_h	人口比为 10%的最高收入人口的平均收入
最低收入水平 q_l	人口比为 10%的最低收入人口的平均收入
收入水平差 Δq	$\Delta q = q_h - q_l$
收入因子 φ（令 $\varphi=\Delta x_h^*$）	$\varphi = \alpha \Delta q$　　$\alpha = \varphi / \Delta q = \Delta x_h^* / \Delta q$
上下限对比系数 β	$\beta = 1/2$

上述约定中 Δx_l^* 和 Δx_h^* 是根据所给收入分配分组数据 x_{j+1} 计算出的数值，是收入下限与中位数的差值 Δx_l 和收入上限与中位数的差值 Δx_h 的估计值。按照上述约定，以题中附录表二（下表 2）中的数据为例，可以计算上述约定数值。

表7　地区 A 年份一数据表

x_j	x_{j+1}	p_j	L_j
0.00	2228.28	0.10	0.0250
2228.28	3066.03	0.20	0.0673
3066.03	3790.18	0.30	0.1221
3790.18	4519.24	0.40	0.1882
4519.24	5254.75	0.50	0.2663
5254.75	6166.38	0.60	0.3569

6166.38	7273.48	0.70	0.4631
7273.48	8813.52	0.80	0.5901
8813.52	11424.93	0.90	0.7485
11424.93	14171.91	0.95	0.8493

$$q_h = \frac{(1-L_9)N\mu}{(1-p_9)N},$$

$$q_l = \frac{L_1 N\mu}{p_1 N},$$

$$\Delta q = q_h - q_l = \frac{(1-L_9)N\mu}{(1-p_9)N} - \frac{L_1 N\mu}{p_1 N}$$

以上公式中，N 表示总人口数量，在计算中是可以约掉的，μ 表示平均收入，表中在最后已经给出，p_j、L_j 的数值也可以在表中查到，这样，经过计算，

$$q_h = \frac{(1-0.7485)\times 6281.34}{1-0.9} = 15797.57,$$

$$q_l = \frac{0.025\times 6281.34}{0.1} = 1570.34,$$

$$\Delta q = 15797.5701 - 1570.335 = 14227.24$$

下面来叙述确定收入因子的方法，也就是确定收入因子中参数 α 的方法。不妨取收入的中位数 $m = x_{5+1} = 5254.75$，假设

$$\Delta x_h^* = x_{8+1} - x_{5+1} = 8813.52 - 5254.75 = 3558.77$$

Δx_l^* 在这基础上乘以一个上下限对比系数 β，β 的范围应该满足 $\beta \in (0,1)$，这在收入分配的密度函数 $f(x)$ 上也可以看出来，当收入 $x > m$ 时，人口比例要少于收入 $x < m$ 时的人口比例，要使中位数 m 左右的人口比例大致相当，则 $\Delta x_h > \Delta x_l$ 是必然的，反映在函数上是:

$$F(x_h) - F(m) = F(m) - F(x_l) \Rightarrow \Delta x_h > \Delta x_l$$

(记对应的分布函数为 $F(x)$，则 $p = F(x)$ 表示收入低于或等于 x 的人口比例)。不妨设 Δx_h^* 作为收入因子，这样 $\alpha = \frac{\Delta x_h^*}{\Delta q} = \frac{3558.77}{14227.24} = 0.25$，收入因子 $\varphi = 0.25\Delta q$，则：

$$\Delta x_h \approx \varphi = 0.25\Delta q = 0.25\times 14227.24 = 3556.81$$

根据 $\beta = 1/2$，也就是说 $\Delta x_l \approx 1/2\,\varphi = 0.5\times 0.25\Delta q = 1778.41$ 这是只要提供中位数 $m = 5254.75$，就可以得到：

$$x_l = m - \Delta x_l = 5254.75 - 1778.41 = 3476.34,$$
$$x_h = m + \Delta x_h = 5254.75 + 3556.81 = 8811.56$$

这样中等收入人口的收入上下限就可以确定了。将 AB 两地各年份的的数据代入到 x_l、x_h 中：

$$x_l = m - \Delta x_l = m - \beta\alpha\Delta q$$
$$x_h = m + \Delta x_h = m + \alpha\Delta q$$

在第 5 节中，我们将使用这种改进的收入空间法来研究 AB 两地在不同年份的中等收入人口变化情况。用这种“中位数调整法”可以算出中等收入的上下限。

4.2 基于人口空间法的改进

“人口空间法”，指的是选择 $F(m)=1/2$ 邻近的一个范围为中等收入人口，例如取范围 $p_1=20\%$ 到 $p_2=80\%$，这样，中等收入人口比例就已经取定为 60%。再用此 60%的人口所拥有的收入占总收入的比例来描述中等收入人口的状态。

传统的人口空间法的缺陷在于，当出现一种两极分化的情况时，例如，某地的收入分配是[1000,3000]上的均匀分布，这时中位收入是 m=2000。此时，中间 60%人口拥有总收入的 60%，中等收入界定范围是 14000 到 26000。到收入分配发生变化，变成了[0,4000]上的均匀分布，这时收入范围拉大了，低端人口收入下降了，高端收入人口收入增加了，直观上两极分化扩大了，也即中等收入人口应该是下降了，但按第二种方法，中间 60%的人口拥有的总收入比例仍是 60%。

传统的人口空间法缺点在当出现两极分化现象时，对于中等收入人口的分析与直观上不符合，中等收入人口所占的比例 60%是达不到的，应小于 60%。而下降的人口比例大小，应是与原中等收入人口比例 p（即 60%）在时间变化的过程中所拥有的收入占总收入的比例 $L(p)$ 的改变量相关，即：原 60%中等收入人口占有的财富比例变化值越大，相应的中等收入人口比例的变化值就越大，中等收入人口财富比例变化值 Δw 与中等收入人口比例变化值 Δp 之间有一个呈正比的系数，用公式来表示：

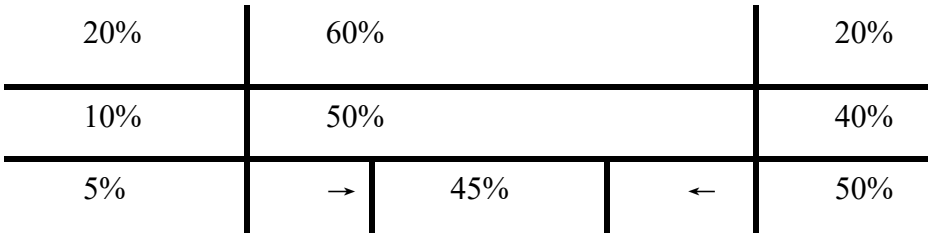
$$\Delta p = \delta \times \Delta w$$

也可以采用另外一种思路，以低收入人口所占有的财富比例变化的绝对值 Δw_l 乘以一个系数 δ_l 作为低收入人口比例的增长值 $\Delta p_1 = \delta_l \times \Delta w_l$，以高收入人口所占有的财富比例变化的绝对值 Δw_h 乘以一个系数 δ_h 作为高收入人口比例的增长值 $\Delta p_2 = \delta_h \times \Delta w_h$，这样中等收入人口比例变化值 Δp 可以表示为：

$$\Delta p = \delta_l \times \Delta w_l + \delta_h \times \Delta w_h$$

当然，中等收入人口的比例并非固定是 60%，也可能是其他数值，这种改进的人口空间法旨在解决在各年中 p_1 与 p_2 取不同的值时，纵向比较各年中等收入人口与收入的变动，给出了解决 p_1 与 p_2 变化的方法。

举个例子来说明所提出的方法，如示意图所示，低收入人口比例 20%，所占的财富比例为 10%，中等收人人口比例 60%，所占财富比例为 50%，高收入人口比例 20%，所占的财富比例为 40%，经过若干年后，假设出现两极分化现象，若仍以低、中、高收入人口比例为 20%：60%：20%来计算财富比例的话，其财富比例由 10%：50%：40%变化为 5%：45%：50%，意味着中等收入人口中有人变得贫穷，有人变得富有，总人数和中等收入人口比例是减少的，低收入和高收入人口比例是增加的，在表中的反应是 p_1 右移， p_2 左移，计算具体移动比例的方法已在上文中提到。



在第 5 部分中，我们将分析改进的人口空间法的效果。

5　模型的实证分析（A,B 地区）

在这一部分中，我们通过分析 A,B 两地的数据了验证基于收入空间法和人口空间法改进效果，同时对各地区、各年份的中等收入范围，中等收入人口数量，以及变化趋势等情况进行了描述。

5.1 用收入空间法分析 A,B 两地的中等收入人口情况

对于各地区中等收入的范围的界定，我们依据问题二中提到的改进的收入空间法来进行确定，确定规则如下公式所示：

$$\begin{aligned} x_l &= m - \Delta x_l = m - \beta\alpha\Delta q \\ x_h &= m + \Delta x_h = m + \alpha\Delta q \end{aligned} \qquad 5\text{-}1\text{-}1$$

根据以上规则，确定了各地区各年份的中等收入人口的收入下限和上限

表8　中等收入上下界

	地区 A 年份一	地区 A 年份二	地区 B 年份一	地区 B 年份二
下限 x_l	3476.34	4915.96	11484.08	14254.99
上限 x_h	8811.56	12533.58	22298.65	28933.87

在问题一的分析中可以看到，我们自己建立的模型一在均方误差和平均绝对误差等参数上明显优于其他洛伦兹模型，因此我们选择该模型进行曲线拟合。

曲线模型如下所示：

$$L(p)=p^{\alpha}[1-L_{\lambda}(p)(1-p)^{\beta}]^{\gamma} \tag{5-1-2}$$

已知了 L(p)与 p 的关系式，由以下公式推导可得到 p 与 x 的关系表达式，即 $F(x)$ 与 x 的函数关系。

$$\frac{dL(p)}{dx}=\frac{dL(p)}{dp}\frac{dp}{dx}=\frac{dL(p)}{dp}\frac{dF(x)}{dx}=\frac{dL(p)}{dp}f(x) \tag{5-1-3}$$

$$\begin{aligned} &\text{又}\because L(p)=\frac{1}{\mu}\int_0^x tf(t)\mathrm{d}t \\ &\therefore \frac{dL(p)}{dx}=\frac{xf(x)}{\mu} \\ &\therefore \frac{xf(x)}{\mu}=\frac{dL(p)}{dp}f(x) \end{aligned} \tag{5-1-4}$$

$$\therefore \frac{dL(p)}{dp}=L'(p)=\frac{x}{\mu},\ x=\mu L'(p) \tag{5-1-5}$$

对于（5-5）式，由于复杂度较高，不能够求出其反函数，在已知 x 的情况下， 我们使用牛顿插值法解方程得到对应的 p_H 和 p_L，根据中等收入界限的上限 x_H 和下限 x_L 来确定中等收入人口分布。

根据上述四种不同的数值确定洛伦兹曲线模型的参数，通过 matlab 内置的函数 lsqcurvefit 获得各组数据的拟合参数，参数如下表所示：

表9　参数表

	α	λ	β	υ
地区 A 年份一	1.6570	-15.6163	0.5645	0.4094
地区 A 年份二	1.6654	-18.0298	0.5915	0.4418
地区 B 年份一	1.4123	-18.9604	0.5543	0.3838
地区 B 年份二	1.4236	-18.3801	0.6729	0.3753

四组数据的洛伦兹曲线以及 x-p 曲线如下所示，同时求得 x-F(x)关系曲线。

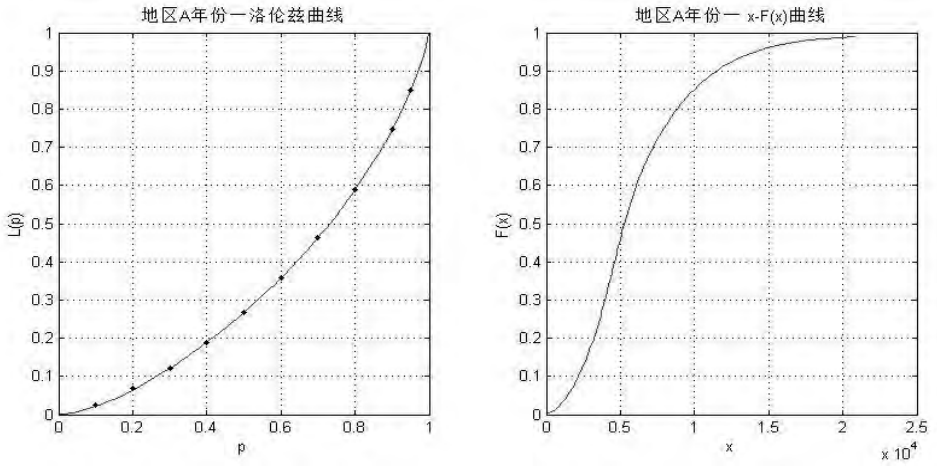


图17　地区 A 年份一洛伦兹曲线及 x-F(x)关系曲线图

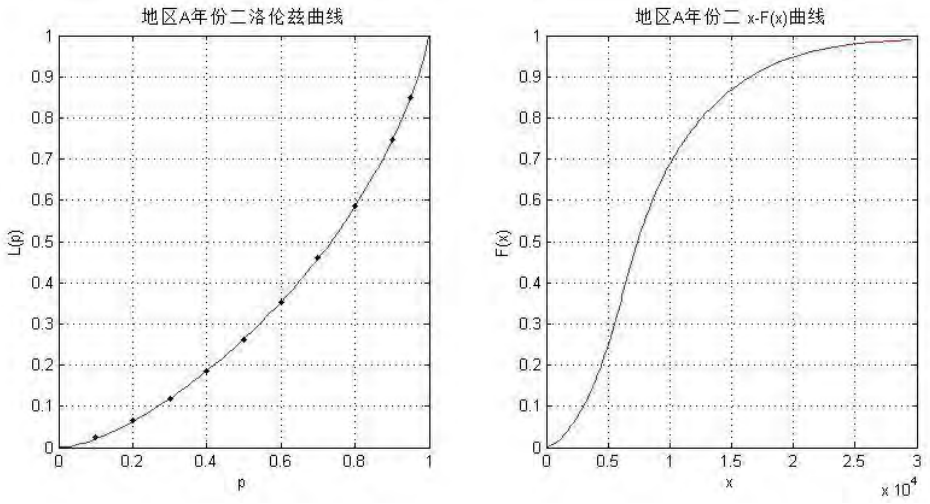


图18　地区 A 年份二洛伦兹曲线及 x-F(x)关系曲线图

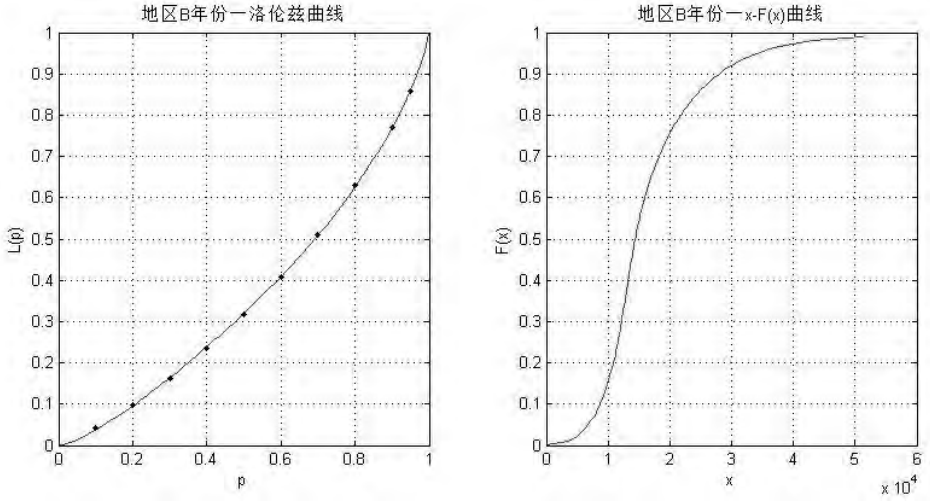


图19　地区 B 年份一洛伦兹曲线及 x-F(x)关系曲线图

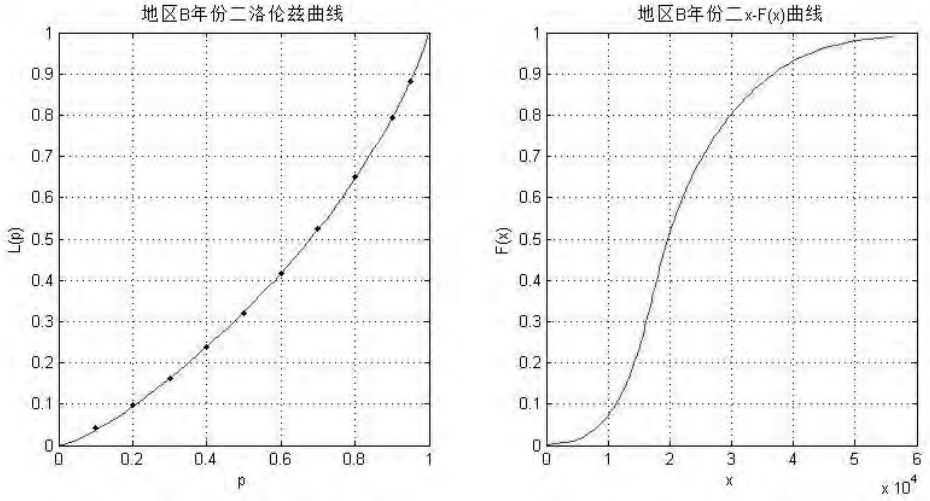


图20　地区 B 年份二洛伦兹曲线及 x-F(x)关系曲线图

由表 1 提供的四组 x_l 和 x_h 值，将值分别带入式（1），通过牛顿插值法，在已知 x 的情况下，解函数方程，求得四组数据对应的 p 值如下

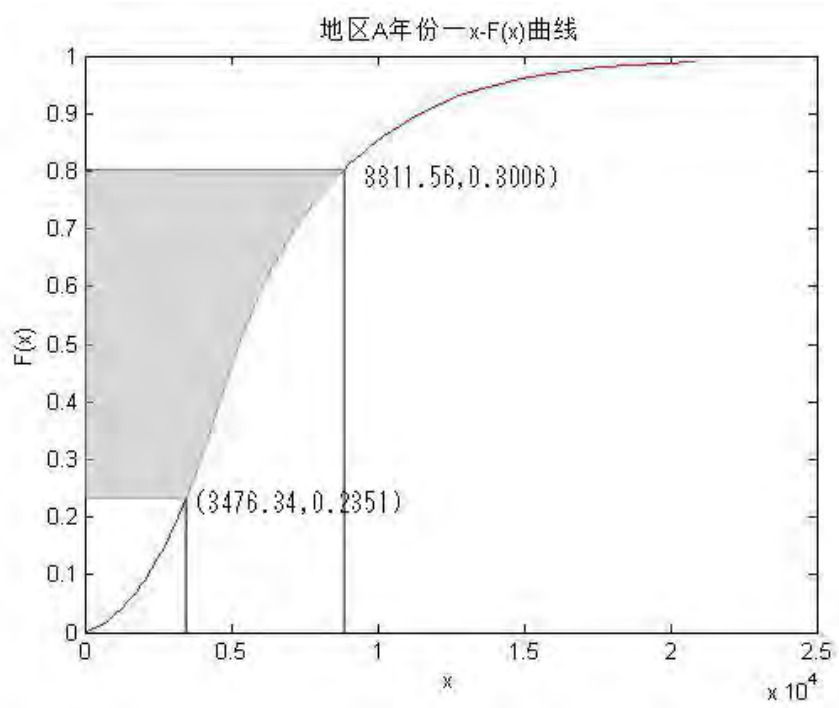


图21　x-F(x)曲线关系图

表10 中等收入 p 值

	地区 A 年份一	地区 A 年份二	地区 B 年份一	地区 B 年份二
下限 p_L	0.2351	0.2422	0.2428	0.2013
上限 P_H	0.8006	0.7977	0.8115	0.7833
人口比例	0.5645	0.5555	0.5687	0.5820

表11 中等收入人口状态 S

	地区 A 年份一	地区 A 年份二	地区 B 年份一	地区 B 年份二
$L(p_L)$	0.0828	0.0852	0.1238	0.0951
$L(P_H)$	0.5900	0.5825	0.6414	0.6244
收入比重	0.5072	0.4973	0.5176	0.5283

根据题意，中等收入人口的状态定义为 $S = L(p_H) - L(p_L)$，这个状态值的含义就是中等收入人口所占的收入占总收入的比重。根据上表给出的各组数据的 p_L 和 p_H，我们可以得到 A 在年份一的中等收入人口状态为 0.5072 ，在年份二中等收入人口状态为 0.4973 ，由此我们可以发现地区 A 中等收入人口的总财富值所占的比重降低了。从表 3 中反映出来的数据我们可以看出，地区 A 的中等

收入人口的比重由 0.5645 变为 0.5555，地区 A 的中等收入人口比重降低了 0.009，同时，中等收入人口的总收入比重也降低了 0.0009。之所以会出现这样的下降，依我们分析来看，是因为 A 地从年份一到年份二贫富差距稍有增大，导致中等收入群体向两端分散。

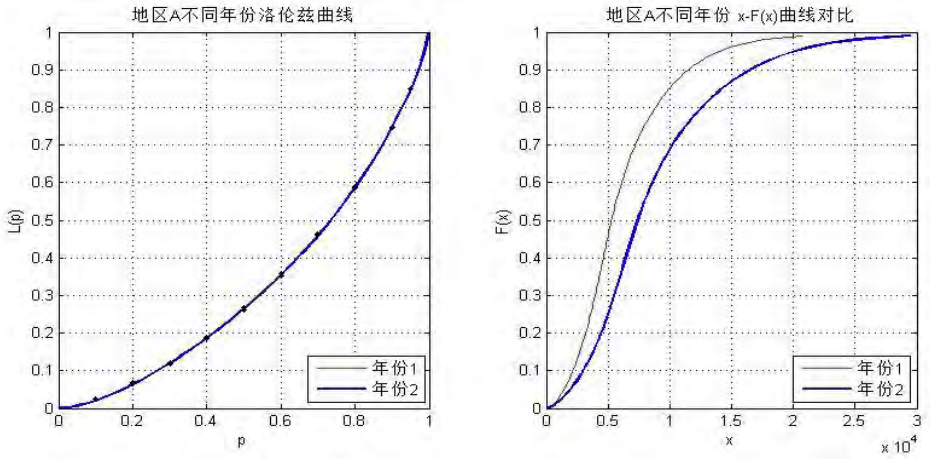


图22　地区 A 不同年份对比

从图 5 中 f(x)-x 的关系曲线图来看，由图 5 我们可以看出，地区 A 在年份一中的 $x \sim F(x)$ 曲线较年份二的更加陡直，这说明经过一定时间的积累，经济发展、收入增加等因素导致所有人口的收入都右移。年份二中 $F(x)$ 左侧区域的面积（即蓝色曲线左侧面积）明显大于年份一中 $F(x)$ 左侧区域的面积（即红色曲线左侧面积），说明年份二的平均收入较年份一有较大增长。

地区 B 在年份一时中等收入人口所占的比重为 0.5687，到年份二时中等收入人口比重上升到 0.5820， p_L 从 0.2428 下降到 0.2013，中等收入群体向下扩充较大，这主要是因为我们设计界定标准是基于收入空间的，从年份一到年份二，地区 B 的平均收入增加了 5280.07，按照我们的标准，有一部分低收入群体，收入增加，进入了中等收入群体的范畴。 p_H 从 0.8115 降低到了 0.7833，高收入群体的比重增加，有一部分中等收入的人，进入高收入者的行列。

增加的收入相对于中等偏上的来说其意义小于低收入者，也就是说进入中等收入人的数量大于进入高收入者的数量，因此总的来说进入中等收入范围的人是增加的。从数据上我们可以看出，中等收入人口的比重增加了 0.0153。中等收入人口所占的收入比重由 0.5176 增长到 0.5283,增加了 0.0117,中等收入人口所占总收入比重稍有增加，这是因为进入中等收入范围的人和进入高收入范围人的收入水平不对等，前者的人均收入低于后者，但是前者有数量优势，因此总的来看，中等收入群体的总收入比重略有上升。

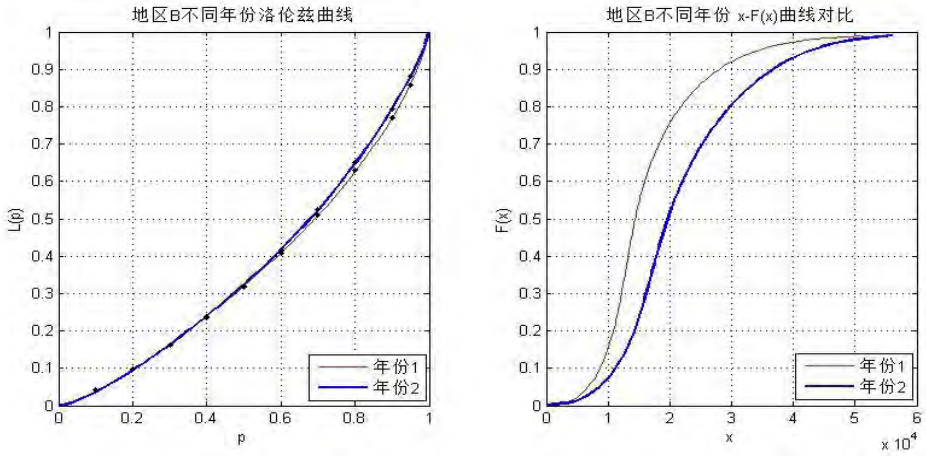


图23　地区 B 不同年份对比

从图 6 中地区 B 不同年份对比图可以看到，年份一的洛伦兹曲线比年份二的略微弯曲一点，这表示年份二的贫富差距比年份一有所降低。从右边的图我们可以看出，年份二的蓝色线条在年份一的红色线条左边，反应了年份二的平均收入明显高于年份一。

由地区 A 年份一和年份二的数据我们也可以看到，年份二与年份一相比，低收入群体的总的收入占比下降，高收入者的总收入占比上升，也就是说，从年份一到年份二，地区 A 两级分化的现象变得严重了，这就导致中等收入群体的比重下降，从我们给出的定量分析结果来看，也验证了这种猜想。

问题（2）横向比较地区 A 和地区 B 在相同年份的中等收入人口、收入情况，

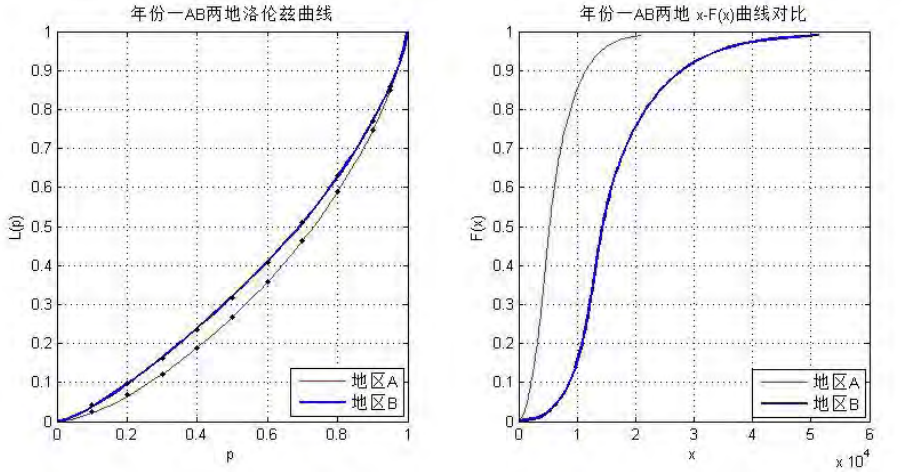


图24　AB 两年年份一曲线比较

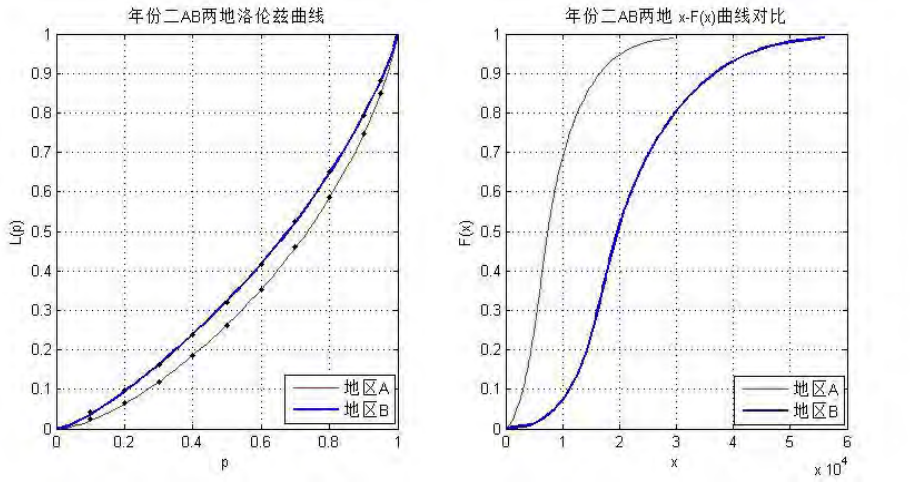


图25　AB 两地年份二曲线比较

相同的一点是地区 B 的收入曲线弯曲程度很明显的小于地区 A 的洛伦兹曲线，地区 B 的基尼系数小于地区 A，这说明地区 B 贫富差距较小，的收入均匀程度要好于地区 A。从平均收入水平上来看，在图 3(b)和图 4(b)中，地区 B 中 $F(x)$ 左侧区域的面积（即蓝色曲线左侧面积）明显大于地区 A 中 $F(x)$ 左侧区域的面积（即红色曲线左侧面积），说明地区 B 的平均收入远大于地区 A 的平均收入，这从表二-表五所给出的数据中也能得到体现。

另外从绝对收入水平上看，在图 3(b)和图 4(b)中，表示地区 B 的蓝色曲线整体都在表示地区 A 的红色曲线的右侧，这说明在同一时间段内，地区 B 的收入水平都要好于地区 A。

5.2 实例分析改进的人口空间法

以地区 A 为例，通过上一小节中提到的改进收入空间法，得到年份之一的中等人口收入区间 (x_l, x_h) 为(3476.34, 8811.56)和中等收入人口比例的区间 (p_1, p_2) 为(0.2351, 0.8006)，中等收入人口所占财富比例的区间 $\left(L(p_1), L(p_2)\right)$ 为(0.0828,0.5900)。

以上述人口空间改进法,将年份之一的中等收入人口比例的区间 (p_1, p_2) 对应到年份之二的洛伦兹曲线中去，得到年分之二中等收入人口所占财富比例的区间 $\left(L'(p_1), L'(p_2)\right)$ 为(0.0807, 0.5913)，这样就可以对比年分之一中等收入人口所占财富比例的区间 $\left(L(p_1), L(p_2)\right)$ 计算出财富比例的变化值：

$$\Delta w_l = L'(p_1) - L(p_1) = 0.0828 - 0.0807 = 0.0021,$$
$$\Delta w_h = L'(p_2) - L(p_2) = 0.5900 - 0.5913 = -0.0013,$$

分别加权系数 $\delta_l = 3.4$ 和系数 $\delta_h = 2.2$，得到中等收入人口比例在年份之一和之二的变化值 Δp，可以表示为：

$$\Delta p_1 = \delta_l \times \Delta w_l = 3.4 \times 0.0021 = 7.14 \times 10^{-3}$$

$$\Delta p_2 = \delta_h \times \Delta w_h = 2.2 \times 0.0013 = 2.86 \times 10^{-3}$$

$$\Delta p = \delta_l \times \Delta w_l + \delta_h \times \Delta w_h = 3.4 \times 0.0021 + 2.2 \times 0.0013$$

这样可以得到年分之二的中等收入人口比例的区间的估算值

$$p_1' = p_1 + \Delta p_1 = 0.2351 + 7.14 \times 10^{-3} = 0.2423$$

$$p_1' = p_1 + \Delta p_1 = 0.8006 - 2.86 \times 10^{-3} = 0.7978$$

对比之前改进收入空间法算出的实际值(0.2422, 0.7977)，相差不大，可以说是非常接近，说明改进的人口空间法可信度较高。

6　中等收入人口测算方法及其经济学意义

6.1 中等收入人口的定义

所谓中等收入，从模糊意义上讲是指介于高等收入和低等收入之间的一种收入水平状态。从收入的绝对量上看，也可以看作是收入水平处于一定地域范围内全体居民某个时期平均收入水平上下的收入。

但中等收入水平只是一个相对概念，而不是绝对的平均水平。中等收入并不是现有收入的简单算术平均，而是相对于高收入和低收入的不同水平而言，处于中间层次、接近于平均和中等偏上、符合全面建设小康社会要求的收入水平。这是因为，如果社会收入结构是“橄榄”型或“哑铃”型，那中等收入必定是社会平均收入水平;而如果社会收入结构表现为“金字塔”型或“倒金字塔”型，社会平均收入水平就必然低于或高于所界定的中等收入水平。

6.2 中等收入人口定义的原理

我们已经知道，洛伦兹曲线作为一种重要的经济分析手段，其横轴表示人口比例，纵轴表示总收入比例，如图 13 所示。我们从本文中大量使用的洛伦兹曲线中得到灵感，构建一个反映社会财富分布平等性的指数，定义曲线 $L(p)$ 下方的面积与 $L(p)$ 上方的面积的比值为财富分布平等指数

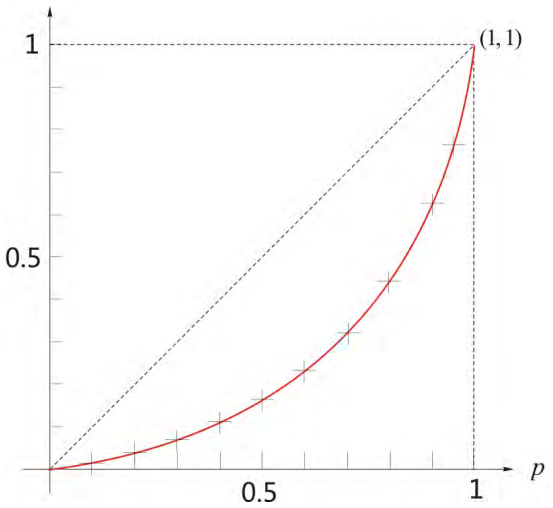


图26　洛伦兹曲线

于是财富分布平等指数定义为

$$\varepsilon = \frac{\int_0^1 L(p)dp}{1-\int_0^1 L(p)dp}$$

根据财富分布平等指数的定义，可以得到$\varepsilon \in (0, 1)$，当财富分布平等指数上升时，反映在洛伦兹曲线上，曲线的弯曲程度下降，中等收入人口比例上升。其中45°线可以理解为平等收入线，当洛伦兹曲线接近45°线时，$\varepsilon \to 1$，任何低收入端人口比例为p的人口拥有的总收入比例也是p，从而必定是完全平等的收入分配。$\varepsilon \to 0$时，说明洛伦兹曲线弯曲程度达到最大，低收入端人口几乎不占据社会财富，高收入端人口几乎占据了社会的全部财富，意味着贫富差距达到一个难以想象的程度，当然，这是不可能达到的，任何国家和地区都会在财富分布平等指数低到一个警戒值的时候采取必要的宏观调控措施，从而扭转贫富差距的拉大。

我们提出的财富分布平等指数应满足以下三条公理：

- 单调性：中等收入人口比例增加时，财富分布平等指数增加；
- 转移性：当低收入人口所占的财富比例增加，或者高收入人口所占的财富比例减少，财富分布平等指数都应该增加；
- 有界性：财富分布平等指数应是有一个取值范围。

6.3测算方法及经济学意义

为了通过国际比较判断中国收入不平等状况，利用世界银行《2006年世界发展指标》公布的世界122个国家基尼系数和“五分法”收入分配数据，绘制财富分布平等指数与中等收入人口比例相关图。

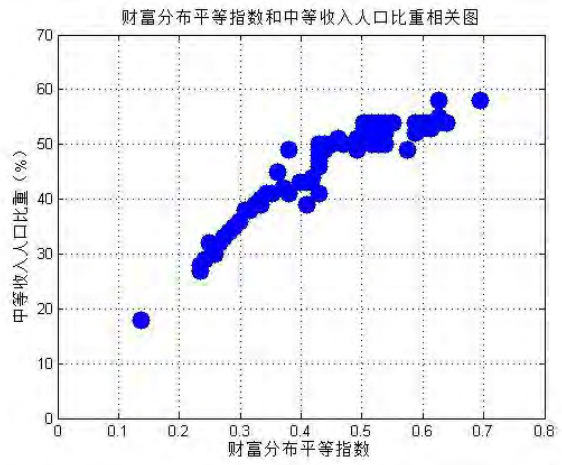


图27　财富分布平的和中等收入人口比重相关图

从图中结果结果表明：全球 122 个国家中，财富分布平等指数 $\varepsilon \le 0.25$ 的国家有 7 个，占全部国家数的 5.7% ； $0.25 \le \varepsilon < 0.35$ 的国家有 19 个，占全部国家的 15.6%；$0.35 \le \varepsilon < 0.45$ 的国家有 32 个，占全部国家数的 26.2%；$0.45 \le \varepsilon < 0.55$ 的国家有 48 个，占全部国家数的 39.3%；$0.55 \le \varepsilon < 0.65$ 的国家有 16 个，占全部国家的 13.1%，财富分布平等指数处在 0.65 以上的国家只有 1 个，占比为 0.8%。财富分布平等指数 ε 处在 0.45~0.55 之间的国家最多，其次是处在 0.35~0.45 之间的国家，它们合占全部国家数的 65.6%。由此可见，世界上绝大多数国家的财富分布平等指数 ε 处在 0.35~0.55 之间，中国的财富分布平等指数也处在这一区间内。其中财富分布平等指数值低于中国的国家共有 34 个，占全部国家数的 27.9%。在财富分布平等指数值低于中国的国家中，既有低收入国家，又有中等国家，但没有一个高收入国家。

财富分布平等指数和中等收入人口比重之间具有明显的正向相关关系。即财富分布平等指数越大，中等收入人口比重越大；财富分布平等指数越小，中等收入人口比重越小。

对世界 122 个国家财富分布平等指数和中等收入人口比重的相关散点，建立一元线性回归模型 $y = a + bx$，得结果如下：

$$R_{mid} = 0.1378 + 0.7703\varepsilon$$

由上述回归方程，我们可知：如果财富分布平等指数保持在 0.35~0.45 之间，则中等收入人口比重应大致在 43%~50% 之间。要保持一国中等收入人口比重不低于 40%（社会稳定所要求的底线），要求财富分布平等指数不低于 0.33（见下表）

表12 财富分布平等指数和中等收入人口比重的对应关系

财富分布平等指数	0.1	0.2	0.3	0.4	0.5	0.6	0.7
中等收入比重（%）	19.78	29.87	38.40	45.72	52.05	57.60	62.50

财富分布平等指数和居民收入分配结构变化是各国政府、企业、居民普遍关注的问题。尤其对于像中国这样由传统计划经济向现代化市场经济转轨的国家，这类问题更有深入研究的必要。人们通常把财富分布平等指数 0.45 作为贫富差

距拉大会导致社会不稳定的警戒线，通过分析表明，财富分布平等指数与中等收入比重之间存在某种正向相关关系：财富分布平等指数越小，中等收入比重越小；而财富分布平等指数越大，中等收入比重则越大。

在收入分配的密度函数 $f(x)$ 上可大致看出，财富分布平等指数低于 0.33 时，收入人口分布结构不合理的“金字塔形”，而当财富分布平等指数高于 0.55 时，收入人口分布结构呈两端小、中间大的“纺锤型”或“橄榄型”，构建“橄榄型”财富结构成为今后我国收入分配改革一项必然的政策选择。

7 总结

本文中我们根据已有基础模型结构，提出了两种洛伦兹曲线模型，并建立了一种多项式模型，结构较简单，而且计算比较方便；在求解参数的过程中，通过 matlab 中的 lsqcurvefit 函数计算，并创新地联想到信号检测与估计理论中噪声概念，提出一种新的求参方法。

对比王祖祥教授提到的十种洛伦兹曲线模型，用均方误差（MSE）、平均绝对误差（MAE）和最大绝对误差（MAS）三种方法来确定每一种模型的拟合精度，然后确定哪一种模型能够更好描述国民收入在国民之间分配的格局，通过对比，我们构建的模型在拟合精度上并不差，洛伦兹曲线模型一的精度甚至要优于部分已有模型。

中等收入人口界定一直是社会关注的热点。已有的研究中提到的中等收入人口界定的方法有收入空间法和人口空间法，但或多或少都存在这一些缺陷，例如收入空间法存在中等收入的范围的任意性，人口空间法在发生两级分化时就会与经济直观不相符。

针对这两种缺陷，我们分别提出了改进的方法。中位数调整法是对收入空间法的改进，确定了中等收入上下限取值的确定函数，这样对于不同地区、不同年份的中等收入人口及其收入区间都有现成可行的划分方法。改进的人口空间法是通过对中等收入人口比例的上下限做出确定函数的改变，以便纵向比较各年中等收入人口与收入。

居民收入分配结构变化是各国政府、企业、居民普遍关注的问题。尤其对于像中国这样由传统计划经济向现代化市场经济转轨的国家，这类问题更有深入研究的必要。对于中等收入人口的测算方法，我们在本文大量使用的洛伦兹曲线中找到灵感，重新构建一种指数来描述其在社会总人口中所占的比重，定义洛伦兹曲线下方面积与上方面积的比值为财富分布平等指数，借此得到中等收入人口的比重。

目前我国收入人口分布结构呈现不合理的“金字塔形”，构建两端小、中间大的“橄榄型”财富结构将成为今后我国收入分配改革的一项必然政策选择。

参考文献

[1] Chotikapanich, D., D. S.P. Rao, and K.K. Tang, 2007. Estimating income inequality in China using grouped data and the generalized Beta distribution. The Review of Income and Wealth 53, 127-47.

[2] Wang, Z.X., Y-K Ng, and R. Smyth, 2011. A general method for creating Lorenz curves. The Review of Income and Wealth 57, 561-582.

[3] Foster, J.E. and M.C. Wolfson, 2009. Polarization and the decline of the middle class: Canada and the U.S. Journal of Economic Inequality 8, 247-273.

[4] 庄健, 2007. 基尼系数和中等收入群体比重的关联性分析. 《数量经济技术经济研究》2007 年第四期.

[5] 顾纪瑞, 2005. 界定中等收入群体的概念_方法和标准之比较. 《现代经济探讨》2005 年第十期.

[6] 纪玉山, 2005. 中等收入者比重的扩大及橄榄型财富结构的达致. 《社会科学研究》2005 年第二期.

[7] Wang, Z.X. and R. Smyth, 2013. A hybrid method for creating Lorenz curves with an application to measuring world income inequality.

[8] Sen, A., 1976. Poverty: An ordinal approach to measurement. Econometrica 44, 219-232.